# WOULD ACCOUNTABILITY BASED ON TEACHER VALUE ADDED BE SMART POLICY? AN EXAMINATION OF THE STATISTICAL PROPERTIES AND POLICY ALTERNATIVES

**Douglas N. Harris**
Educational Policy Studies
University of
Wisconsin–Madison
1025 W. Johnson Street #575K
Madison, WI 53706
dnharris3@wisc.edu

**Abstract**

Annual student testing may make it possible to measure the contributions to student achievement made by individual teachers. But would these “teacher value-added” measures help to improve student achievement? I consider the statistical validity, purposes, and costs of teacher value-added policies. Many of the key assumptions of teacher value added are rejected by empirical evidence. However, the assumption violations may not be severe, and value-added measures still seem to contain useful information. I also compare teacher value-added accountability with three main policy alternatives: teacher credentials, school value-added accountability, and formative uses of test data. I argue that using teacher value-added measures is likely to increase student achievement more efficiently than a teacher credentials-only strategy but may not be the most cost-effective policy overall. Resolving this issue will require a new research and policy agenda that goes beyond analysis of assumptions and statistical properties and focuses on the effects of actual policy alternatives.

## 1. INTRODUCTION

Annual standardized student testing is a pervasive, and probably permanent, piece of the U.S. K–12 education system. But this has not resolved the age-old issue about whether and how policy makers should use the test results in accountability systems. During the 1970s, many states expanded student testing and adopted minimum competency exams that students had to pass to graduate from high school. In the 1980s, they added school report cards to the accountability mix, reporting point-in-time snapshots of average school achievement. This trend toward test-based school-level accountability accelerated in the 1990s with state policies such as school grades, reconstitution, takeovers, and other incentives (Harris and Herrington 2006). The No Child Left Behind (NCLB) Act appears to have cemented school-level, test-based accountability as a key lever in the national education strategy, but this is a broad strategy and leaves the door open for a wide variety of policies regarding the use of standardized test scores. One new policy option that has intrigued players on all sides of the education debate is accountability based on how much "value added" teachers and schools contribute to student achievement.

A fundamental problem in holding schools accountable for student achievement is that, as economists put it, education is jointly produced by schools, families, and communities (Hanushek 1979). Students' socioeconomic status is arguably the strongest predictor of their educational outcomes, an observation dating at least as far back as Coleman (1966). There is also strong evidence that this correlation reflects a causal relationship. The achievement levels of black kindergarteners are half a standard deviation below the levels of white kindergartners (Fryer and Levitt 2004). Because such differences occur before students enter school, they must be due to family, community, and other factors outside of school control.[1] Rothstein (2004) describes the myriad ways that family and community factors influence student learning. It is therefore no surprise that a school serving white students from middle- and high-income families is eighty-nine times more likely than a high-minority, high-poverty school to be among a state's top third on achievement tests (Harris 2007).

These facts pose difficulties for school-level accountability systems whose expressed goals are to measure and reward school performance. If school performance measures substantially reflect nonschool contributors to student success, as is the case with NCLB and typical state school report cards, then genuine improvements in school performance will not show up in higher performance measures. This leaves schools with weak incentives to improve

[1] Lee and Burkham (2002) also provide extensive evidence on these "starting gate inequalities" using the same database as Fryer and Levitt (2004).

and leads to perverse incentives to "cream" the most socioeconomically advantaged students (Harris 2007) and push out low performers (Figlio 2005). In this sense, school accountability based on such misleading performance measures is unfair not only to schools but to students as well.

Value-added modeling has drawn wide interest in recent years as a way to solve this problem and isolate the contribution of schools. The term *value added* comes from the economics literature and refers to the contribution of inputs to outputs in a production process. In the education literature, the term is used to describe analyses using longitudinal student-level test score data to study the educational input-output relationship, including especially the effects of individual teachers (and schools) on student achievement. The basic logic is simple: if each student's achievement is measured every year, then in trying to determine each teacher's performance we can take into account where students started at the beginning of each year and therefore indirectly account for the family and community factors that also contribute to achievement. Value added differs from typical school report cards and NCLB, which utilize snapshots of student achievement at a point in time and do not account for where students start; as a result, they fail to accurately measure school performance.[2]

Value added can also be used to measure the performance of teachers and hold them accountable as individuals. The U.S. education system has a long history with one form of teacher accountability—teacher merit pay—dating to the early part of the last century (Murnane and Cohen 1986). Showing renewed interest, many districts such as Denver and states such as Florida are again experimenting with merit pay, using student achievement as a key component of the performance measures on which merit and compensation bonuses are paid. Political leaders at the federal level have also taken up the idea. Some school districts are funding their merit pay plans with the federal Teacher Incentive Fund (TIF), and President Obama has proposed additional funding for TIF in his first budget request to Congress.[3] New efforts in Congress are poised to create additional federal supports.[4] An alternative proposed by Gordon, Kane, and Staiger (2006) and considered by some school districts

[2] Report cards in some states, such as Florida and Kentucky, do include achievement growth as part of the school performance measures, but these are exceptions. Under NCLB, the U.S. Department of Education has now allowed fifteen states to pilot "growth models," but this name is misleading, and schools are still primarily held accountable for achievement levels rather than growth.

[3] Full disclosure: The author was a member of the Technical Advisory Committee for the Teacher Incentive Fund.

[4] Congressman George Miller (D-CA), a leader on education issues in the U.S. House of Representatives, recently proposed a program in which school districts could apply for funds to provide additional compensation to teachers in low-performing schools if teachers demonstrated performance on measures such as student test scores. For more recent discussions of the background and evidence on teacher merit pay, see Figlio and Kenny (2007) and Podgursky and Springer (2007).

such as New York City, is to use teacher value added as a "key component" for teacher tenure decisions (p. 2). More than just adding accountability, these compensation and tenure policies take aim at some of the oldest and firmly established teacher-related policy traditions—nearly guaranteed job security and compensation based on credentials (i.e., the single salary schedule).

Even many advocates of test-based accountability, however, acknowledge that measuring teacher contributions to student test scores is difficult. Teacher value added is intended to address these difficulties, but, as I describe in the next section, several assumptions have to hold in order to interpret value-added measures as true (causal) contributions to student achievement. Many of the assumptions seem unrealistic and have been rejected, most recently in studies commissioned as part of the National Conference on Value-Added Modeling. Most of these papers are included in the present volume and are summarized and interpreted in what follows below.

No performance measure or accountability policy is perfect, of course, and it may be that the assumptions are not violated too severely and that teacher value-added accountability is better than the alternatives. I compare teacher value-added accountability with alternatives such as rewards for teacher credentials, accountability based on school value added, and formative uses of standardized tests as alternatives to teacher value-added accountability, using the policy validity framework outlined by Harris (2008a). The first element of the framework, *statistical validity*, refers to the relationship between any teacher quality measure and the construct it attempts to measure; there is widespread agreement that one critical construct is teachers' contributions to student achievement, and that is the central focus here. The second element of the framework is the *purpose*. There are two main purposes with regard to student achievement scores. First, accountability, by definition, involves creating signals or summative assessments of effectiveness—that is, determining who is performing well as the basis for incentives. But knowing who is performing well does not provide a path to improvement or, borrowing the language of teacher educators, it is not "formative." Formative and summative assessments are interrelated in that any path to improvement may be of little use unless teachers have incentives to improve; likewise, providing incentives for teachers to improve without a path to improvement may do little to drive performance. The third and final element of the framework is cost. While several researchers emphasize the fact that the costs of education programs are just as important as their effects (Harris 2009; Levin and McEwan 2001), there remains little evidence on the cost effectiveness of education programs (Levin 1991; Rice 2002), especially teacher quality initiatives.

After discussing the policy validity of teacher value-added accountability, I provide brief discussions of the three policy alternatives. Based on a comparison of the alternatives using the policy validity framework, I conclude that teacher value-added accountability has real potential, but determining whether that potential can be realized requires studying the costs and effects of actual accountability policies and alternatives. Studies of assumptions and statistical properties of the measures can take us only so far. What is needed is an entirely new research and policy agenda.

## 2. POLICY VALIDITY OF TEACHER VALUE ADDED

### Value Added and Its Assumptions

From a statistical standpoint, isolating the impact of schools from that of families and communities is a problem of nonrandom assignment of teachers to students. If all teachers had the same chance of being assigned to any given student, and if we had enough observations on each teacher, we could draw conclusions about each teacher's effectiveness simply by looking at the end-of-year test results, and complex statistical adjustments would be entirely unnecessary. But there is ample evidence of nonrandom assignment—for example, that the most disadvantaged students are assigned to the least qualified teachers (Clotfelter, Ladd, and Vigdor 2005) and "tracked" into classrooms within schools that have other disadvantaged students (Gamoran 1986; Oakes 1985; Ogbu 2003). The potential attraction of value added is that it may allow us to indirectly account for family and community factors even when these types of nonrandom assignments arise.

While the popularity of the term value added is relatively new, the ideas and their application to education date at least as far back to Hanushek (1979) and Boardman and Murnane (1979). For more recent discussions that account for the rapid advances in student data, see, for example, Todd and Wolpin (2003), Harris and Sass (2005), and Harris (2010). Below I briefly describe some of the key assumptions of the models and recent evidence about their validity, including evidence from many of the articles in this volume.

Researchers have considered a wide variety of approaches to estimating teacher value added. While the purpose here is to discuss the assumptions and statistical properties that are shared by most or all of these methods, a few comments are warranted about the specific model specifications. First, while the focus here is primarily on the methods, terminology, and past studies by economists, the main issues discussed are common to noneconomics models and to the variations within the economics approach.[5]

Second, I focus the discussion on value-added models (VAMs) that compare each teacher's performance with broad groups of teachers, including

[5] See Harris and McCaffrey (2009) for a comparison of economics-based models with those from statistics in general and educational statistics.

those in other schools, and avoid models that compare teachers only to their colleagues within their respective schools. This is important because even most of the strongest advocates of test-based accountability express some concern about pitting teachers against one another within the same school and potentially undermining teamwork and collegiality. For that reason, only comparisons of teachers across schools are relevant here; in statistical terms this means omitting school fixed effects. As we will see below, this constraint on the estimation of VAMs has consequences for the statistical validity of the models.

Finally, Harris (2008a) distinguishes value added for accountability (VAM-A) from value added for program (VAM-P) evaluation. Generally speaking, the assumptions required for valid estimates of programs are much less stringent because VAM-A requires obtaining unbiased estimates for every teacher, while estimating the effect of a program on teacher value added can be estimated without bias, as long as the bias in individual teacher effects is unrelated to teachers' participation in the program of interest (e.g., formal teacher education). For this reason, the ideal model for VAM-A, which is the focus here, may differ from the ideal VAM-P model, and each type needs to be judged according to somewhat different standards because each is used to draw different types of conclusions. Following are the assumptions for VAM-A.

***Assumption #1:*** *School administration and teamwork among teachers do not have a significant impact on student achievement.*

The first implication of the need to compare teachers across schools is that it becomes quite difficult to account for the impact of school administration. If we were making within-school comparisons, we might reasonably assume that the impact of school administration affects all teachers relatively equally so there is no need to account for it. But when estimating VAMs for accountability, this approach fails because very few teachers are observed in multiple schools, which would aid in isolating the effect of the teacher from the effect of administrators. This leaves one of two options: (1) measure the quality of school administration directly (e.g., through surveys of teachers and parents) and include these in the VAM; or (2) assume that the impact of administration is small. Information from surveys is rarely, if ever, used in external accountability systems, which limits the practicality of the first option. This leaves the second option, which I label Assumption #1.

A similar problem arises with teacher teamwork. The purpose here is to measure how much each teacher contributes to student achievement, but it is possible, contrary to the assumptions of value added, that teachers contribute to the achievement of students of other teachers—for example, by mentoring.

The only rigorous evidence I am aware of on this point is Harris and Sass (2007b), who find that the number of National Board for Professional Teaching Standards (NBPTS) teachers in a school has no impact on the value added of other teachers within the same schools, but this is far from definitive.[6] It is possible that neither administration nor teamwork plays a significant role; some researchers describe teaching as "loose coupled," meaning that teachers mainly work on their own in their classrooms, making it difficult for anyone else to have a significant impact on what they do or how well they do it.

Overall, the evidence on Assumption #1 is thin. If teamwork and administration were important factors affecting teacher effectiveness, they would be difficult to account for in VAMs and could introduce bias.

***Assumption #2***: *Controlling for previous achievement levels is sufficient to account for the impact of past school resources.*

Education is a cumulative process. The educational resources students receive early in life affect their academic success later in life. But as a practical matter, it is impossible to explicitly measure the whole range of resources students receive at any given time, let alone in past years. It may be possible to account for resources indirectly, however, because the effects of all resources should be reflected in subsequent achievement. This means that when trying to explain why students reached achievement level $A$ at time $t$ ($A_t$), we can account for past school resources by controlling for achievement in the previous time period ($A_{t-1}$). The effects of all school resources experienced up to time $t-1$ should be reflected in $A_{t-1}$.

Notice that Assumption #2 involves only "school resources." Controlling for past achievement also accounts for family and community factors, but only under the assumption that students are assigned to teachers based solely on their previous achievement and not on unobserved student characteristics that may also be related to students' subsequent achievement. This is implausible. For example, Feng (2005) finds that students are assigned to teachers partly based on students' discipline problems, which are generally unobserved. Also, Harris and Sass (2005) and McCaffrey et al. (2009) show that the findings regarding teacher value added are quite different when relying solely on Assumption #2 to account for student differences. Because of the centrality of family and community factors, researchers have developed VAMs that do not require such restrictive assumptions (see Assumption #3 below).

There is some question about the degree to which past resources (including teachers) influence student achievement relative to more recent resources,

[6] There are other studies of mentoring, but in those cases teachers have formal mentoring roles. Here we are interested in all the ways teachers influence one another.

often referred to as "decay" or "fade-out."[7] At one extreme, past schooling resources (e.g., last year's teacher) may have very little impact on current achievement. At the other extreme, past schooling resources could be just as important as current ones in affecting current achievement. Kane and Staiger (2008) show that individual teacher effects decay by 50 percent or more per year. That is, the impact of having a good teacher does not seem to last. As Rothstein (2009) points out, this could be because the variation in teacher value added is driven by differences in instruction that have only ephemeral impacts, for example, how much teachers teach to the test. Another possible explanation is that the content of achievement tests is somewhat independent across years. To take an extreme example, suppose students need not understand any of the academic content covered on the third-grade test in order to learn the academic content on the fourth-grade test. In that case, we would not expect the third-grade teacher's contribution to the third-grade achievement score to have any impact at all on the fourth-grade test. Whatever the explanation, the apparently high rate of decay is not an assumption of VAMs and does not necessarily pose a problem in terms of the validity or bias of value-added estimates. Harris and Sass (2005) find that the impact of school resources and programs (in VAM-P models) is relatively insensitive to any decay assumption that might be imposed, though this might not be the case in VAM-A models.

While there is some debate about exactly how to incorporate past achievement, the fact that past achievement is a very strong predictor of current and future achievement suggests that accounting for it is very important and that doing so does indeed account for the vast majority of past schooling inputs.

***Assumption #3*****:** *Students' contributions to their own achievement can be measured with student fixed effects that account for the nonrandom assignment of students to teachers ("static selection").*

As noted above, it would be unrealistic to assume that students are assigned to teachers based on measurable qualities only. To address this, economists typically include student fixed effects in their VAMs, which represent the (conditional) average rate of achievement growth over all the years they are in the database. They are conditional in the sense that these control for students'

[7] Economics-based value-added models typically assume that decay is geometric; that is, the effects of past teachers on current achievement declines at a constant annual rate. The rate of decay can be estimated directly by including lagged achievement as an independent variable. Value-added models that use the change in score as the dependent variable assume zero decay. The empirically estimated rate of decay depends on other aspects of the model specification. For example, the rate of decay in a model with student fixed effects is likely to be lower because, in the absence of student fixed effects, lagged achievement reflects both average achievement and the year-specific deviation. With student fixed effects, lagged achievement reflects only the latter.

average school resources and other factors that might influence student learning in any given year that are largely outside teachers' control and that might influence student learning.[8] Rothstein (2009) describes this as the "static selection" assumption. This does not preclude changes over time in students' propensities to make learning gains, but it does mean that any time-varying propensities are randomly distributed among teachers. Otherwise, there is what Rothstein (2009) calls "dynamic selection," which may introduce bias into the teacher value-added measures even when student fixed effects are included. The specific nature of the assumed static selection depends somewhat on the model specification.

The static selection assumption with student fixed effects is almost certainly more realistic than the alternative of no selection based on unobservable qualities that is required in the absence of student fixed effects (see Assumption #2 and Harris and Sass 2007a), which explains why economists typically include student fixed effects in their models. But the matter is still not completely settled. Kane and Staiger (2008) report on an experiment involving seventy-eight classrooms in the Los Angeles School District. The researchers solicited school principals willing to randomly assign teachers to classrooms within their schools. The researchers then compared the value added measured before the experiment to those calculated on the basis for random assignment, which, as long as the random assignment was carried out with fidelity, cannot be driven by systematic assignment of students to teachers. Specifically, they regressed mean end-of-year test scores on previous value added. A coefficient of one on the value-added variable would seem to suggest that value added is a perfect predictor of teacher contributions when random assignment is used. For some value-added specifications they indeed find coefficients close to one.[9]

In addition, Rothstein (2009) tests the dynamic selection assumption by considering whether the teacher assignment in any given year predicts *past* achievement growth. While we would expect the *current* teacher to affect *current* achievement, a current teacher cannot change what has already happened—or "rewrite history"—and will appear to do so only when students are nonrandomly assigned. Rothstein estimates VAMs with student fixed effects and

[8] As discussed by Harris (2010), the fixed student contribution is often called *innate ability* by economists and is akin to what psychologists consider general intelligence, or *g*. The more general term, *fixed student contribution*, is used here because it is virtually impossible with education data sets to estimate anything like innate ability. No data sets include measures of student abilities at birth or, in their absence, sufficiently measure family and other environmental factors well enough to distinguish innate from environmental differences.

[9] Kane and Staiger (2008) do not report results from the common specification with achievement gains and student fixed effects because they found that the student fixed effects were jointly insignificant. They do estimate a model with achievement *levels* and student fixed effects but find that this performs poorly compared with several alternatives. The correct specification is still not a settled issue, but the Kane and Staiger results are compelling.

indeed finds that current teacher assignment does predict past student achievement and therefore rejects the static selection assumption.

It is worth considering how violations of the static selection assumption might arise in practice. The most obvious explanation, and the example commonly given to explain the above findings, is that school principals "track" students and do not randomly assign teachers to tracks. Monk (1987) finds that most school principals randomly or evenly distribute students in elementary grades, apparently because principals want to even out the workload among teachers. But he also finds that some principals try to match students to teachers who have skills particularly well suited to students' needs. This violates the static selection assumption and may explain why Rothstein (2008) finds that the assignment of future teachers predicts past student achievement gains.[10]

Because we are focused on VAMs in which teachers are compared across schools, another form of potential selection bias involves the nonrandom assignment of teachers to schools. Principals cannot randomly select teachers from the entire population of potential teachers, or even from the entire pool within their respective school districts. Instead they can choose only from among the teachers who apply for jobs in their respective schools. There is ample evidence of nonrandom assignment of teachers to schools, and that assignment is correlated with factors (teacher experience, etc.) that are sometimes related to teacher value added (Clotfelter, Ladd, and Vigdor 2005). It is unclear whether this makes the selection bias problem any worse. One possible way to avoid potential problems here is to compare teachers across *similar* schools. This approach is used at the school level (that is, comparing schools with similar student demographics) in England (see Ray, McCormack, and Evans 2009). The same approach also turns out to be helpful in solving another problem discussed below.

**Assumption #4:** *A one-point increase in test scores represents the same amount of learning regardless of the students' initial level of achievement or the test year.*

[10] Rothstein (2008) also discusses the issue of principal assignment decisions, writing that "it requires in effect that principals decide on classroom assignments for the remainder of a child's career on the day that child begins kindergarten" (pp. 12–13). This statement unintentionally makes the assumption seem less realistic than it is. As noted above, the assumption of value-added models is satisfied under the "even distribution" assumption, even if the decisions about even distribution are made "dynamically" such that principals take into account time-varying information about students. It would therefore be more accurate to say, in the context of within-school comparisons of teachers, that the models assume that some principals randomly assign students and the remaining principals make decisions about each year's track based solely on the previous year's track, without making use of any new information. This still seems implausible, but a little less so than Rothstein's formulation. Also note that Rothstein's evidence seems to reject even the weaker assumption.

Value-added models are, at a basic level, models of student achievement. Therefore it is unsurprising that value added requires strong assumptions about the measurement of student achievement. Specifically, it is assumed that a one-point change in the score is the same on every point on the test scale—that is, the test is interval scaled. Even the psychometricians who are responsible for test scaling shy away from making this assumption in the strict sense.

Some adjustments can be made in the value-added analysis to account for the scale problems. For example, some researchers add grade-by-year fixed effects, which essentially equalizes the mean achievement (or achievement gain depending on the specification) across grades and years. This approach is sufficient as long as the scaling problems are limited to differences in the scale over time and/or across grades that affect only the average gain, but the problem is almost certainly more complicated than that. An arguably better, and increasingly common, approach is to "normalize" all the test scores to a mean of zero and a standard deviation of one, based on the standard deviation of the respective grades and years. Such adjustments probably improve the validity of the estimates but because the nature of the test scaling problems is still essentially unknown, they are really ad hoc solutions.[11]

Ballou (2009) argues that the assumptions of traditional scaling techniques, based on item response theory (IRT), are inherently difficult to test. Further, even the plausibility of the resulting test scales from these methods is questionable, and other reasonable approaches yield quite different measures of achievement gain. Ballou describes an alternative non-IRT method of measuring student progress, requiring less restrictive assumptions, in which students are ranked based on their achievement gains and then teacher value added is calculated based on changes in student rankings rather than the gains themselves.[12] He finds that the rankings of teachers on their value added often vary dramatically between the traditional IRT approach and the student rank-order approach, even though cases can be made for each. Briggs and Weeks (2009) also examine sensitivity to test scaling and find less sensitivity than Ballou, but this is likely due to: (1) the narrower range of assumptions that they consider (all fall within the IRT paradigm); and (2) the fact that they focus on school value added rather than teacher value added.

The interval scale assumption requires that a one-point increase in the test score means the same thing on every part of the test scale, or that the test is "globally" interval scaled. Alternatively, one could impose a less restrictive

[11] This is not the only assumption required regarding the properties of the student achievement tests. For example, there is also an implicit assumption that the content of the tests is constant over time.

[12] The advantage of ordinal scales is that they require less restrictive assumptions, although they do throw out potentially useful information.

assumption that tests are "locally" interval scaled, meaning that one point has to mean the same thing only over a fairly narrow range. In other words, the global interval scale assumption requires that a one-point increase for a student in the 10th percentile means the same thing as a student at the 90th percentile, whereas the locally interval scaled assumption requires only that the one-point increase means the same thing for students at the 30th and 70th percentiles, which is probably more realistic. The local interval scale assumption therefore could be operationalized by comparing each teacher only with others whose students have similar initial test score levels—that is, who start off on the same part of the test scale. Ballou (2009) calls this "binning," and while it almost certainly improves matters, he argues that it may not solve the problem.

***Assumption #5:*** *Teachers are equally effective with all types of students.*

The fact that students and teachers are not randomly assigned has already been established. One potential problem that arises from this is that some teachers might be assigned to students who are less likely to make achievement gains. Even if the VAMs succeed in accounting for this, teachers may vary in how much they contribute to learning of different types of students.

To see the problem more clearly, suppose that some teachers were effective with low-achievement students and other teachers were effective with high-achievement students. Further, suppose that all teachers were assigned only to students with whom they were most effective and that in such a situation all teachers appear equally effective in their value-added scores. Now suppose instead that some teachers were "mis-assigned" by principals to students with whom they were ineffective, and as a result their value-added scores decrease. These same teachers who had been judged effective will now appear ineffective simply because of the assignment process. This is problematic because teachers cannot control to which students they are assigned, and it would be difficult to argue that these mis-assigned teachers are really less effective than the others.

The above example is an extreme case, intended to illustrate the potential problem created for value added if teachers are not equally effective with all students. Lockwood and McCaffrey (2009) conclude that differential effects explain less than 10 percent of the variation in overall teacher effects. Therefore what seems like a potential issue in theory may not be significant in practice.

The above five assumptions do not represent an exhaustive list of assumptions that apply to all value-added models, though they are arguably the ones that are considered to be potentially most problematic.[13] Other

[13] Another assumption is that student test data are missing at random. The data requirements for value added are significant, and those data will be missing for a large portion of the students due

assumptions vary depending on the model specification. Harris and Sass (2005) test a variety of these assumptions. It is also important to point out that these assumptions may be interrelated so that violating one assumption might compound or offset the impact of violations in other assumptions. Research at present is mainly focused on testing individual assumptions, which is often quite complicated by itself.

### Statistical Properties of Teacher Value Added

It is possible that all the assumptions of VAMs are violated but that the violations are not so severe that they have a practical impact on value-added measures and the associated accountability rewards and sanctions. Conversely, all the assumptions might hold but the models might still not have the statistical properties necessary for particular types of policy uses. This section explores other empirical findings regarding value added that are relevant to understanding their usefulness for accountability.

#### *Teacher Value Added Is Positively Correlated with Other Measures of Teacher Effectiveness*

Teacher value added can be viewed as an objective measure of teacher effectiveness in the sense that the method of calculating it is the same for all teachers and is not filtered through the subjective preferences and beliefs of a supervisor or other evaluator. There is a long history of research studying the relationship between subjective and objective measures of worker productivity as well as the implications of this relationship for employment contracts.

Teacher value added appears to be positively correlated with principals' confidential assessments of teachers (Harris and Sass 2007c; Jacob and Lefgren 2005). After adjustments for measurement error, these correlations are in the 0.3–0.5 range depending on the model specification (Jacob and Lefgren 2005). Principals in both studies were asked about teacher performance broadly defined, and teachers' contributions to student achievement are likely to be only one part of how principals define performance. Indeed, there is clear evidence that principals considered factors other than student achievement in making their confidential assessments (Harris et al. 2008; Harris and Sass 2007c).[14]

to absenteeism, mobility across schools, and data processing errors. Missing data do not bias the results so long as they are missing at random, though missing data significantly diminish the reliability of the estimates. This is a strong assumption and is especially likely to be a problem in high-poverty schools where absenteeism and mobility are high and test-taking rates are lower. It is therefore a significant question whether valid value-added estimates can be made in schools with high mobility.

[14] In addition to asking for their overall subjective assessments, the authors in these two studies asked principals how well teachers contributed to student achievement so they could determine how much weight principals gave to student achievement in their overall assessments. These alternative measures correlated at 0.7, suggesting that student achievement is probably the main objective of

So even if principals knew exactly how much teachers contributed to student achievement, we would not expect the confidential principal assessment to equal teacher value added. For the same reason, we should not view the correlation between value added and confidential principal evaluations as simple validity checks.

The positive correlation works both ways, of course, so another possible response to this evidence is that we should just use principals' assessments instead of value added. However, there is a significant difference between asking principals to give their assessments confidentially to a researcher versus making a public assessment that would influence the career or compensation of a teacher. There is good reason to think that public assessments of teachers by principals, or anyone else who has a personal relationship with the teacher, would be inflated because the principals would want to avoid discord and hurt feelings, and certain teachers would receive preferential treatment that is unrelated to any objective notion of performance. Value-added measures are not subject to this type of inflation and bias.

### *Value-Added Measures Have Been Replicated in a Randomized Control Trial*

Recall that Kane and Staiger (2008), in their study of the Los Angeles School District, were able to nearly replicate random assignment-based estimates of teacher effectiveness with nonexperimental value-added estimates. Conducting an experiment of this sort is inherently difficult, which makes Kane and Staiger's work especially impressive. However, there is one limitation that makes it difficult to view this as a validation of teacher value-added measures. Specifically, it is unclear how principals were assigning teachers before the experiment took place. If they were assigning teachers in essentially random ways, then the "experiment" is really no different from what was already happening and their results could not be interpreted as evidence in support of value added.[15] On the other hand, if principals were tracking students and nonrandomly assigning teachers to different types of teachers, the results here are significant and reinforce the potential of teacher value added.

Based on these findings—that teacher value added is correlated with principal evaluations and has been replicated in a random assignment experiment—the news on value added reinforces the potential use of value added for accountability. This is not the case with the following two findings.

these principals but also that other outcomes such as motivation and socialization likely explain the modest size of the correlation between the two measures. For this reason, the comparison of principal evaluations of teachers with teacher value-added measures cannot be viewed as a validity check per se, but it does suggest that value-added measures provide useful information.

[15] Their findings regarding value-added specifications would still be valid even if principals had been randomly assigning teachers and students to begin with. Each specification makes different assumptions, as the earlier discussion highlights, and the goal is to get as close to the experimental estimates as possible.

### *Teacher Value-Added Scores Are Imprecise*

A prerequisite for any performance measure to be useful is that different teachers obtain different scores. Sanders and Horn (1998) and Rivkin, Hanushek, and Kain (2005), for example, find considerable differences between the most and least effective teachers based on value-added results, and this is partly why there has been so much interest in using teacher value added for accountability—it raises the possibility of being able to weed out the low performers and reward and attract more high performers. However, a substantial share of this variation could be due to different forms of statistical error, which not only exaggerates the amount of actual variation in true teacher performance but reduces the useful information in the measures. The issue here is one of reliability.

Kane and Staiger (2001, 2002) provide one of the best and most well-known discussions of the types of errors in value added. Using data from North Carolina, they concluded that only about half the variation in grade-level achievement gains is due to "persistent" differences between schools—that is, to differences that could plausibly be attributed to factors under the control of the schools. This is noteworthy given that their analysis was conducted at the grade level where classrooms are grouped together and where the amount of imprecision is therefore likely to be better than for individual teachers. Reinforcing this point, they showed that the persistent component of grade-level gains was considerably smaller in schools with fewer students. Other researchers have shown that teacher value-added scores are imprecise enough that, by the usual standards of statistical significance, it is possible to clearly distinguish only very low value-added teachers from very high value-added teachers (Jacob and Lefgren 2005). This is a problem for policies that intend to make high-stakes decisions based on the measures, except perhaps if those decisions pertain only to rewards for very high performers and punishments (e.g., rejection of tenure) for very low performers. But even this may be problematic because it means some truly average teachers will be rewarded or punished unjustifiably.

Some of the "other non-persistent" variation identified by Kane and Staiger (2001, 2002) is driven by measurement error. Boyd et al. (2008) explain how to account for measurement error using methods often called "shrinkage estimators" and show that the impact of measurement error is greater in value-added models than in cross-sectional models, for the simple reason that value-added measures are based on changes in student achievement, and changes in any measure are statistically "noisier" than measures at a point in time. While accounting for measurement error reduces the observed variation in teacher value-added scores, it certainly does not eliminate it, which means that there are still meaningful differences in teacher performance. This improves matters, but even the adjusted teacher value-added measures do not appear to reach typical standards of reliability (more on this below). This is important for accountability because it means that the judgments made about teacher performance based on value added may be incorrect fairly often.

### *Individual Teacher Value Added Is Unstable over Time*

Intuitively, we would expect that the actual effectiveness of each teacher changes little from year to year. Teachers might gradually improve over time, but it is unlikely that they will jump from the bottom to the top of the performance distribution. It is even less likely that true teacher rankings on value added should drop significantly over a short period of time, except perhaps in cases such as divorce or other significant change in teachers' family status or health.

Some of the earliest evidence on this topic, however, suggests that teacher value added is much more unstable than this intuition would suggest. Koedel and Betts (2007) found that only 35 percent of teachers who were ranked in the top fifth of teachers on teacher value added in one year were still ranked in the top fifth in the subsequent year. This suggests that 65 percent of high-performing teachers actually got worse relative to their peers over a short period of time—some dramatically worse. Stability appears somewhat higher in studies by Aaronson, Barrow, and Sander (2007) and Ballou (2005), but this may be due solely to the fact that these two studies divided teachers into only four groups instead of five groups as in Koedel and Betts, making it less likely that changes in groups would be observed. Overall, these results are remarkably similar across studies.

A substantial portion of this instability is due to small samples and measurement error. McCaffrey et al. (2009) show that while reliability coefficients of unadjusted single-year teacher value-added measures are in the 0.4–0.8 range, stability increases by 40–60 percent when aggregating data across two years and an additional 18–23 percent when adding a third year. This is important because it is likely that accountability policies calling for the use of teacher value-added measures would include requirements that many years of data be used for each teacher. In addition, 30–40 percent of the variance in teacher value added is due to measurement error, which can also be accounted for with shrinkage corrections. This suggests that the reliability of teacher value added in practical applications might approach the typical research standards (a reliability coefficient of 0.8), though reliability will likely remain a weak point.

Like Koedel and Betts (2007), McCaffrey et al. (2009) also find that stability greatly improves when student and school effects are omitted. This is a predictable result because there is mobility of teachers across schools that changes the basis of comparison each year. There is much less variation in the entire pool of teachers in a school district or state and therefore less change in the population with whom each teacher is being compared. The fact that stability is greater without school fixed effects is advantageous because such effects need to be excluded to prevent pitting teachers against one another within schools.

The remaining instability may be due to genuine changes in teacher effectiveness over time, which value-added measures are intended to capture, or to violations in the assumptions. For example, as noted earlier, VAMs assume that accounting for past achievement is sufficient to account for past resources. If instead, as Rothstein (2009) suggests, teachers are assigned based on unobserved time-varying student characteristics, and these unobserved characteristics (or the process of nonrandom assignment) change over time, this might generate "false" instability. Also, if each teacher's value added did vary considerably across student groups, then year-to-year changes in assignment of students to teachers, combined with differential impacts, would reflect true changes in teacher value added that are larger than the above intuition alone might suggest. I revisit some of these statistical properties later because many other policy approaches require similar assumptions and have similar statistical properties.

In short, teacher value added seems promising because the measures are correlated with principals' confidential assessments of teachers, and they have been validated in some sense within a randomized control trial. On the other hand, the measures are imprecise and therefore bounce around from year to year in ways that do not appear to reflect actual teacher performance.

### Purposes and Costs of Teacher Value Added

The policy validity framework outlined by Harris (2008a) includes not only statistical validity, a topic covered in the previous two sections, but the purposes of the measures. This means that the above discussion of statistical properties has little meaning without specifying the types of conclusions one wishes to draw. The implicit assumption above is that teacher value added is intended to create signals that (potentially) provide information about which teachers contribute the most to student achievement. Therefore teacher value added might tell us how well teachers are performing overall but tell us nothing about how they might improve.

The choice of purpose is also relevant to the third piece of the framework—cost. Here I consider both the standard opportunity cost definition as well as budgetary costs. The costs of making teacher value-added calculations, or any other statistical adjustments to student tests scores, are small. If we assume the tests will be administered with or without the value added, the only additional cost is limited to hiring some expert staff or consultants who are knowledgeable about value added to make the calculations. An additional cost is explaining the meaning of the calculations to educators—this is far from trivial because the measures can otherwise be misunderstood or not taken seriously. Also, since the purpose here is to hold teachers accountable, it is arguably also necessary to include not just the costs of value-added measures themselves but also the costs of the related accountability mechanisms. Harris et al. (2008) show that the budgetary costs of teacher merit pay plans can be quite high unless they can be implemented by shifting existing teacher compensation funds. This seems somewhat unlikely because it would require cutting salaries of the lower-performing teachers, which would be politically unpalatable. Other accountability policies based on teacher value added, such as use in tenure decisions, would require few resources of any kind.

The fact that there are many ways in which teacher quality measures might be used in policy makes it difficult to generalize about policy validity. However, the framework does suggest that if the goal of education is to raise student achievement, teacher value added is a plausibly cost-effective option: it focuses on the outcome of interest (achievement), has some desirable statistical properties for creating signals of effectiveness, and has policy uses that involve little cost. I return to this again below because policy validity also requires comparisons with policy alternatives.

## 3. POLICY VALIDITY OF TEACHER CREDENTIALS

To make any fair judgment about teacher value added for teacher accountability, it is necessary to compare it with other policy options for improving the quality of instruction. As Harris (2009) points out, the number of possible options (or what he calls "far substitutes") is quite large, so I focus on several options that are often discussed in the context of the value-added debate (teacher credentials) as well as others that represent alternative uses of student test scores.

One of the most widespread policy traditions for improving teaching is to reward credentials—experience, certification, and formal education. Unlike teacher value added, credentials potentially serve two purposes. If teachers with more or better credentials are more effective, then credentials are signals of effectiveness. Also, some credentials—especially formal education, on-the-job training, and professional development—are potential paths to improvement. Some forms of training may improve teacher effectiveness even if they are weak signals of effectiveness (Harris 2008a).

### Teacher Credentials: Effects and Signals

It is important to distinguish between two types of teacher credentials: those that vary over time and those that are fixed. Teacher personality is an example of a relatively fixed characteristic and is often measured in teacher selection instruments such as the Teacher Perceiver. Undergraduate education is another example because very few teachers are in the classroom full time before they have their degrees. Other forms of teacher education, such as graduate training and professional development, change over time. The distinction between fixed and time-varying credentials is important partly because it highlights what can be learned about the policy validity of different types of measures. For a characteristic that is fixed in nature, or one that might vary but is only measured at a single point in time in a particular data set (e.g., undergraduate education), we can easily learn whether the measure is a good signal of teacher effectiveness, but it is much more difficult to determine whether the quality of the signal is due to some unmeasured characteristic of teachers that is correlated with the measured one, or whether improving one's standing on the fixed measure actually causes teacher improvement.[16] In contrast, it is easier to determine the causal effects of alterable and time-varying credentials, such as teacher experience and professional development, because individual teachers can be compared before and after the change takes place. VAM-P models are useful for identifying valid signals of performance and identifying the causal impacts of time-varying teacher credentials. This is true for the same reason that they are useful for accountability: they account for selection bias.[17]

Based on Harris and Sass (2007a), I am aware of twenty-eight studies of the effects of teacher education and experience on teachers' contributions to student achievement, using either the gain score, value-added, or experimental methods. Table 1 summarizes the results from these studies, dividing them into two categories based on the methods used. For reasons explained by Harris and Sass (2007a), as well as above in the discussion of value-added assumptions, the value added and related types of studies are probably more valid than the gain score studies.[18] Note that the numbers in the table add up to a number considerably larger than twenty-eight because many of the studies have estimates of more than one teacher credential. Only one of the studies (Harris and Sass 2007a) includes all the teacher credentials listed in table 1.

[16] In some ways the distinction between fixed and time-varying credentials reiterates the distinction made earlier between signals and improvement, but there is a subtle difference. Signaling and improvement have to do with the function that the measures serve, whereas the fixed versus time-varying distinction has to do with the type of data that are available to the researcher. Credentials that are fixed in the data can be used to study only the usefulness of the measures as teacher quality signals, whereas time-varying credentials can be used to study both signaling and improvement. Some examples of this distinction are provided later in this article.

[17] One form of selection bias—the nonrandom assignment of students to teachers—has been discussed above. Harris and Sass (2007a) describe a second form that involves the nonrandom assignment of teachers to credentials.

[18] Table 1 includes the studies together with a very small number of related studies that address the issues of nonrandom selection using data in which students and teachers are actually or apparently randomly assigned to one another (these address only one form of selection bias).

**Table 1.** Summary Results of Value-Added and Earlier Related Studies

| Teacher Credentials | Gain Score Studies | | Value Added or Related | |
|---|---|---|---|---|
| | **Significant, Positive** | **Insignificant, Negative** | **Significant, Positive** | **Insignificant, Negative** |
| Undergraduate | 5 | 4 | 1 | 2 |
| Graduate | 3 | 10 | 3 | 6 |
| Professional development | 0 | 1 | 2 | 1 |
| Experience | 7 | 8 | 8 | 1 |
| Test score | 5 | 2 | 1 | 1 |

*Note:* Based on review by Harris and Sass (2007a).

Some studies find a positive and statistically significant relationship between teacher credential and teacher effectiveness, as indicated in the Positive/ Significant category. Other studies find either an insignificant relationship or (rarely) a negative and significant one, indicated by Insignificant/Negative.

Most measures of formal teacher education, especially graduate education, appear unrelated to teacher value added. In the gain scores studies, eight of the twenty-three estimates of the effects of teacher education (undergraduate, graduate, and professional development) suggest that some aspect of teacher education is positively associated with teacher effectiveness. The same finding holds for six of the fifteen value-added or related types of estimates that have studied teacher education. Most of the remaining studies find statistically insignificant associations between education and teacher effectiveness. Harris and Sass (2007a) provide evidence that certain types of teacher professional development (those providing pedagogical content knowledge) lead to improvement in teacher effectiveness.

Teacher experience is consistently positively associated with teacher effectiveness, at least for the first several years. Roughly half of the gain score studies found a positive effect of teacher experience. The effects are overwhelmingly positive in the value-added and related studies, making teacher experience the characteristic that is most clearly related to teacher effectiveness. These results for teacher experience are consistent with evidence on worker experience in other occupations (Harris and Rutledge 2009). This suggests that teachers, as well as other workers, learn not only through formal coursework but also by doing—through their own trial and error.

Teacher test scores are inconsistently associated with teacher value added. The gain score studies in table 1 suggest that teacher test scores are consistently positively related with teacher effectiveness. Only two studies have considered teacher test scores with value added and related methods, but these have yielded more mixed results. Clotfelter, Ladd, and Vigdor (2005) find a positive relationship, whereas Harris and Sass (2007a) find no effect.[19] Research in other occupations, especially complex ones such as teaching, suggests that scores on tests of cognitive ability are positively associated with various measures of job performance (Harris and Rutledge 2009). The tests used in studies of teachers vary considerably in what is being measured—ranging from cognitive ability to teachers' content knowledge and understanding of theories of child learning.

Teacher certification is generally not associated with teacher value added, but evidence on a relatively new and distinctive certification warrants additional attention. NBPTS certification appears to be a moderate signal of teacher effectiveness, with some mixed results across states (Clotfelter, Ladd, and Vigdor 2005; Goldhaber and Anthony 2007; Harris and Sass 2007b). A recent extensive review of this evidence has concluded that National Board teachers have higher value added than others (Hakel, Koenig, and Elliott 2008). NBPTS is an especially interesting credential because it highlights clearly the distinction between the signaling and improvement purposes. The above studies of NBPTS consider not only whether NBPTS is a good signal of value added but whether the process of certification increases teacher value added. While improvement is arguably not the main purpose of NBPTS, it is plausible that such impacts might arise because NBPTS involves over two hundred hours of work by teachers, more than many professional development programs.[20] None of the studies suggest, however, that NBPTS has any impact on value added.

### Costs of Credentials

The most costly teacher quality measure is almost inarguably the master's degree in teacher education, which involves nearly a thousand hours of teacher time spent in class and completing assignments.[21] At $20 per hour, the degree costs at least $20,000 in teacher time alone. This time commitment is five times as long as the time commitment of NBPTS certification and perhaps one hundred times larger than some professional development programs. And these figures ignore the costs of the programs themselves—faculty salaries, university classroom space, and so on. If these were added, the direct costs would only grow.

[19] This may be because Harris and Sass (2007a) controlled for a wide variety of other factors such as coursework. If teacher candidates with greater cognitive ability are more likely to take certain types of college courses, this may make the effect of cognitive ability look smaller than it is.

[20] This calculation was made as follows: suppose that a master's degree requires ten semester-long courses, each of which meets three hours per week for fifteen weeks and requires an equal amount of time outside the classroom: 10 courses × 15 weeks × 6 hours = 900 hours.

[21] Harris and Sass (2007a) report that NBPTS certification requires roughly two hundred hours of work. Professional development programs vary widely.

When the credentials are used as the basis of compensation programs, as is typically the case in public (and most private) schools, the costs just listed may be dwarfed by the budgetary costs of additional salaries. If a teacher with a master's degree earns $3,000 more per year than a teacher without the degree and the teacher stays for twenty years, this could cost the school district $60,000 over the teacher's career—three times more than the costs of teacher time just mentioned.

One of the main reasons for interest in teacher value added is that the credentialing approach is seen as ineffective and costly. The review of evidence above generally reinforces that perception but also highlights the important, and sometimes overlooked, distinction between signals and paths to improvement. Later I show how these purposes, and therefore the potential cost-effectiveness of both teacher value-added accountability and credentials, are interwoven.

## 4. POLICY VALIDITY OF OTHER ACHIEVEMENT-BASED ACCOUNTABILITY POLICIES

While the arguments for teacher value added are often framed in terms of a comparison with the teacher credential strategy, the more obvious alternatives to teacher value added are other uses of student tests scores. In this section, I consider school value added and formative uses of student assessments.[22] This is followed by a comparison of teacher value-added accountability with the credentialing policy and with the other two uses of student test scores.

### School Value Added

The same general method described above for teacher value added can be used to measure school value added and has several advantages, though school value added is easier to estimate for a variety of reasons. First, school administration and teacher teamwork are captured as part of the calculation, so we need no longer make Assumption No. 1. In other words, with school value added, part of the point is to measure and reward both quality instruction and administration and teamwork. Moving from the teacher to the school level is also another way to avoid pitting teachers against one another.

An additional advantage of school value added is that there are roughly ten times as many students per school as per teacher. This goes far toward addressing the imprecision problem with teacher value added. Note also that school value-added measures appear to be less sensitive to violations of the

[22] The words "formative uses of student assessments" reflect the fact that student tests can be designed as "formative assessments," but state standardized tests do not fall into that category. Instead we are talking here about using tests that are designed to be summative but that could be put to formative uses.

assumption regarding test scaling (compare the results in Ballou 2009, who studies teachers, with those of Briggs and Weeks 2009, who study schools).

Another problem with teacher value added is that it can be calculated only for a small percentage of teachers—those who teach for several consecutive years in tested grades and subjects. School value added solves part of this problem, for example, by still incorporating the value added of teachers who teach tested grades for only a year or two. It does not solve other aspects of this problem. For example, gym and music teachers will still contribute little, if anything, to student achievement, and this is no less true with school value added.

There are two important disadvantages of school value added, however. First, school value-added accountability is subject to the free rider problem. If the whole school is rewarded or punished based on school value added, the incentives for effective performance, both within the classroom and in teamwork, may be weak. On the other hand, there is evidence, as noted above, that principals' evaluations of teachers are correlated with teacher value added, and there is anecdotal evidence that "everyone knows" who the high performers are. To the degree that this is true, the school-level incentives could place considerable pressure on principals to attract, hire, develop, and retain high-performing teachers.

School value added would almost certainly be a more accurate measure of school contributions to student achievement than the current federal and state accountability systems that reward only the level of proficiency and therefore do not account for the large role of family and community factors (Toch and Harris 2008).[23] While actual school performance is partly reflected in such measures, the fact that it is confounded with other, perhaps more powerful, forces means that any school effort to improve student test scores is much less likely to show up in higher school grades. Instead, such systems primarily reward schools for

[23] There are some noteworthy differences in the estimation of school and teacher value added. The basic approach to estimating teacher value added rests on the fact that we observe the vast majority of students with multiple teachers. While this process may be nonrandom (as Rothstein 2009 shows), the built-in annual changes in teachers is advantageous for estimating teacher value added. The same cannot be said of school value added. One approach to estimating school value added is to focus on students who switch schools and see how their performance changes. But the very fact that only some students switch automatically creates concern about selection bias (e.g., the types of students who leave may vary across schools), so this is not a wise approach to estimating school value added. Another method would be to make use of the changes that occur as students finish the last grade in a given school and rely on the fact that high schools typically have multiple middle school "feeder" schools and middle schools have multiple feeder elementary schools. Such an approach is more akin to the estimation strategy of teacher value added. In cases in which there is only one feeder school, it would be necessary to rely strictly on the rate of achievement gain compared with other schools and assume that the initial achievement level accounts not only for past school resources but for everything important about the student other than the current teacher. Another issue is that school value-added measures would have to find some way to incorporate school performance in grades that precede testing, typically K–2, as well as grade 3, which in most states and school districts provides the baseline measure for achievement gains in subsequent grades.

attracting students from advantaged backgrounds and pushing out students from less advantaged backgrounds. The so-called "growth models" approved by the U.S. Department of Education in some pilot states do little to fix the problem. In short, these models measure whether students are learning fast enough that they could eventually reach proficiency. This means that schools serving low-performing students are expected to get these students to learn at a *faster* rate than high-performing students. This is unrealistic and creates the same perverse incentives as the proficiency-only model that typifies NCLB.

### Formative Data Uses

Another quite different use of test scores involves giving the student data to teachers by specific test topics (sometimes called "strands"), without calculating teacher or school value added. While state standardized tests are not "formative assessments" in the way this term is typically used, using the tests in this way does constitute a formative use. Knowing exactly where students are performing poorly would allow teachers (and administrators) to target their improvement efforts. School districts are increasingly using state tests this way, some through the adoption of additional quarterly assessments that measure student progress throughout the school year (Burch and Hayes 2007).

The advantage of this approach is that it provides useful information to teachers about how they and their students are doing, information specific enough to help teachers improve. Teacher and school value added, in contrast, provide only signals, which are important for creating incentives but insufficient to drive improvement. This formative use of the data is also potentially inexpensive if it involves providing only state standardized tests data in disaggregated form rather than additional tests such as quarterly assessments. As with other uses of student test scores, providing professional development to teachers to help them interpret and respond to the test results is likely to be an important and more costly part of the process.

## 5. TEACHER VALUE ADDED VERSUS THE ALTERNATIVES

The first comparison of interest is between teacher value-added accountability and teacher credentials. The evidence presented in table 1 is indicative of the widespread perception that teacher credentials neither signal teacher effectiveness nor improve it, and this partly explains why the idea of teacher value added has generated so much interest. If we are interested in maximizing student achievement, then measuring teachers' contributions directly, rather than relying on indirect signals of effectiveness such as credentials, would seem like a better approach.

But the comparison is more difficult than it appears. The first difficulty is that many teacher credentials serve both signaling and improvement purposes.

To the degree that university degrees and certification can be viewed as signals of teacher performance that are explicitly rewarded through hiring decisions and the single salary schedule, it appears that teacher value added stacks up well compared with teacher credentials in the policy validity framework. However, teacher credentials also provide a path to improvement that value added does not. While one could argue that improvement is difficult and that the key to improving teaching is to make sure that the better teachers are hired and retained, it is hard to argue that schools do not also need to create a system and culture of improvement or that training options are unnecessary to facilitate such improvement.

Signals of effectiveness and paths to improvement are also interrelated. Under the present system, the motivation to obtain credentials offers little motivation to genuinely improve. Instead teachers have an incentive to do as little as possible to obtain their credentials. Within the context of a university course or group professional development program, this incentive is hardly conducive to genuine improvement. In *How to Succeed in School without Really Learning,* Labaree (1997) argues that students' efforts to make high marks makes the entire education system worse by focusing student attention on getting good grades rather than learning the material. It is reasonable to expect that this same phenomenon applies to teacher education, especially graduate education, where, at least anecdotally, teachers take university courses mainly because they are required to do so in order to move into school administration or to obtain a higher salary. Thus one reason the credentials may seem largely unrelated to teacher value added (see table 1) is that teachers are getting less out of the credentials than they would if the incentives were set up differently. If teachers sought out credentials on their own in order to improve their performance (e.g., value added), they would not only be more likely to seek out the best credentials but would also be more likely to put forth the kind of effort that would make the credentials useful. Thus the fact that credentials seem unrelated to teacher value added is not a reason to eliminate credentials, but it is a reason to reform the incentive structure that drives them.

Perhaps the more direct comparisons come from other non–mutually exclusive uses of student achievement scores. Since teacher value-added accountability is arguably the most controversial policy under consideration here, let us consider a situation in which intensive school-level accountability is already in place and teachers already have access to strand- or topic-level scores. In this situation, the best-case scenario is that teacher value-added accountability brings all the benefits that its advocates propose, eliminating the free rider problem that remains with school-level incentives. A worst-case scenario, however, is that teacher value-added accountability would reinforce all the negative unintended consequences of the current system, turning education into one large game of teachers pressuring principals to give them the students who they think will yield high teacher value-added scores and teachers instructing students primarily in how to answer particular types of test questions rather than imparting genuine long-term learning. This worst-case scenario is all the more plausible if the value-added measures really have low statistical validity and reflect behaviors that are unrelated to true performance.

A middle ground between these extremes is that teacher value added might turn out to be superfluous if these alternatives were adopted. If school-level incentives already provide significant pressure on teachers to improve and if the achievement data were provided to teachers in a way to facilitate improvement, then they have no additional impact. McCaffrey and Hamilton (2007) provide some evidence that this middle ground is likely. Studying samples of school principals who recently received information about their teachers' value added, they found that most principals did not use the information to change their decision making. The possible impact of teacher value-added accountability is therefore far from clear.

## 6. CONCLUSION

A great deal of attention has been paid recently to the statistical assumptions of VAMs, and many of the most important papers are contained in the present volume. The assumptions about the role of past achievement in affecting current achievement (Assumption No. 2) and the lack of variation in teacher effects across student types (Assumption No. 4) seem least problematic. However, unobserved differences are likely to be important, and it is unclear whether the student fixed effects models, or any other models, really account for them (Assumption No. 3). The test scale is also a problem and will likely remain so because the assumptions underlying the scales are untestable. There is relatively little evidence on how administration and teamwork affect teachers (Assumption No. 1).

The assumptions are important, but even more significant are the statistical properties of the measures. To what degree does teacher value added reflect true differences in teacher performance? Kane and Staiger (2008) find that some value-added models can replicate teacher performance when teachers and students are randomly assigned. There is also evidence that teacher value added is positively correlated with principals' own confidential assessments of teachers (Harris and Sass 2007c; Jacob and Lefgren 2005). This evidence suggests that despite the problematic assumptions, teacher value added still provides useful information about teacher performance. On the other hand, teacher value-added measures are somewhat unreliable, so clear distinctions can be made only between the very highest and very lowest levels of teacher value added by traditional statistical standards. This imprecision partly explains why teacher value added is so unstable over time (Koedel and Betts 2007), although it appears that much of the instability problem can be addressed by using multiple years of data and adjusting for measurement error (McCaffrey et al. 2009).

Some potential policies that would use teacher value added for accountability also stack up well in the policy validity framework when compared with teacher credentials. This is hardly surprising, given that the assumed goal of education here is to raise student achievement. I sometimes use a tennis analogy to make this point. If we wanted to figure out who were the best tennis players, we could carefully observe the backhand technique, footwork, serve percentages, and so on, and from that we could draw conclusions about who is better. Or we could just see who wins the most games. If winning is the goal, trying to incorporate the winning percentage into the performance measurement system is a reasonable thing to do. Measuring the equivalent of the winning percentage is more difficult in education, but the evidence here suggests that it would be worth trying.

How, then, should researchers and policy makers proceed? Given the apparent potential of teacher value added, I recommend that federal and state governments provide funds to encourage local experimentation and learn which policies work in practice (Harris 2008b). Likewise, state and federal governments should avoid putting up legal barriers to experiments with teacher value added, as the New York State Legislature did when New York City Chancellor Joel Klein proposed using student test scores in tenure decisions. Whatever the weaknesses of our decentralized system, the ability to experiment on a small scale is a clear strength and one we should take advantage of. Doing so would be largely a waste of time, however, if such experiments were not accompanied by rigorous evaluation. The federal government has already made this mistake in the TIF grants by imposing only minimal standards on the evaluations.

Ideally, local experimentation would be done through cooperation between local unions and district management. In contrast, some merit pay plans have been forced on teachers, and both sides share the responsibility in most of these cases. Teacher unions are right to call for collaboration, but not as an excuse to maintain the status quo. Likewise, district administrators cannot expect their calls for reform to be embraced when their arguments are accusatory and their proposals ill informed. Not all school districts have the leadership and capacity to lead the way in these policy changes. Experimentation should take place where success is most likely, providing potential examples for others to follow.

One of the possible alternatives to teacher value-added accountability is improved school-level accountability. As I have argued elsewhere (Toch and Harris 2008), NCLB and federal accountability will never reach their full potential and may even be substantially counterproductive if school performance continues to be measured by the percentage of students meeting proficiency. Such measures largely reward schools for who they teach rather than how well they teach, and this does little to provide incentives for real improvement. Further, it makes little sense for state and federal governments to intervene in "failing" schools if they have not correctly identified who is failing. The remedy should match the disease. School value added also solves a political problem because schools, under the current system, have a legitimate excuse to ignore federal accountability. This is ironic given that the present system was motivated by a desire among some advocates to stop the "no excuses" mentality of schools. More important here, an accountability system that starts with school-level value added could improve the effectiveness of teacher-level value-added policies by aligning the entire school performance system. Or it may turn out that teacher value added adds little once a robust school-level accountability system is put in place.

The evidence in this volume is important because it suggests that teacher value added has the potential to improve educational policy and student achievement. It is now time to take this effort to the next stage with a new research and policy agenda focused on putting the idea into practice.

The author was chair of the National Conference on Value-Added Modeling, which included events at UW–Madison, 22–24 April 2008, and the Urban Institute in Washington, DC, 23 May 2008. I wish to thank co-chairs Adam Gamoran and Stephen Raudenbush, program committee members Henry Braun, J. R. Lockwood, Robert Meyer, and Tim Sass, as well as Jane Hannaway, who led the Urban Institute event. For their useful comments on this article, I wish to thank Bob Floden, Adam Gamoran, Drew Gitomer, David Monk, Tim Sass, and participants in the National Conference on Value-Added Modeling, the 2007 Educational Testing Service Invitational Conference in San Francisco, CA, and a seminar at the University of Colorado at Boulder. I also gratefully acknowledge financial support from the Carnegie Corporation, Educational Testing Service, the Joyce Foundation, and the Spencer Foundation.

## REFERENCES

Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics* 25: 95–135.

Ballou, Dale. 2005. Value-added assessment: Lessons from Tennessee. In *Value-added models in education: Theory and applications*, edited by Robert Lissitz, pp. 272–303. Maple Grove, MN: JAI Press.

Ballou, Dale. 2009. Test scaling and value-added measurement. *Education Finance and Policy* 4 (4): 351–83.

Boardman, Anthony E., and Richard J. Murnane. 1979. Using panel data to improve estimates of the determinants of educational achievement. *Sociology of Education* 52: 113–21.

Boyd, Donald, Pam Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2008. Measuring effect sizes: The effect of measurement error. Paper presented at the National Conference on Value-Added Modeling, University of Wisconsin–Madison, April.

Briggs, Derek, and Jonathan P. Weeks. 2009. The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy* 4 (4): 384–414.

Burch, Patricia, and Tracy Hayes. 2007. Accountability for sale: The K–12 testing industry, district contracting, and NCLB. Unpublished paper, University of Wisconsin.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2005. Teacher-student matching and the assessment of teacher effectiveness. Unpublished paper, Duke University.

Coleman, James. 1966. *Equality of educational opportunity*. Report No. OE-38000. Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education.

Feng, Li. 2005. Hire today, gone tomorrow: The determinants of attrition among public school teachers. MPRA Paper No. 589, University Library of Munich.

Figlio, David. 2005. Testing, crime, and punishment. NBER Working Paper No. 11194.

Figlio, David, and Lawrence W. Kenny. 2007. Individual teacher incentives and student performance. *Journal of Public Economics* 91: 901–14.

Fryer, Roland G., and Steven D. Levitt. 2004. Understanding the black-white test score gap in the first two years of school. *Review of Economics and Statistics* 86 (2): 447–64.

Gamoran, Adam. 1986. Instructional and institutional effects of ability grouping. *Sociology of Education* 59: 185–98.

Goldhaber, Dan, and Emily Anthony. 2007. Can teacher quality be effectively assessed? National Board certification as a signal of effective teaching. *Review of Economics and Statistics* 89 (1): 134–50.

Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. Identifying effective teachers using performance on the job. Discussion Paper No. 2006–01, Brookings Institution.

Hakel, Milton, Judith Anderson Koenig, and Stuart W. Elliott, eds. 2008. *Assessing accomplished teaching: Advanced-level certification programs*. Washington, DC: National Research Council, National Academic Press.

Hanushek, Eric A. 1979. Conceptual and empirical issues in estimating educational production function issues. *Journal of Human Resources* 14: 351–88.

Harris, Douglas N. 2007. High flying schools, student disadvantage and the logic of NCLB. *American Journal of Education* 113: 367–94.

Harris, Douglas N. 2008a. The policy uses and "policy validity" of value-added and other teacher quality measures. In *Measurement issues and the assessment for teacher quality*, edited by Drew H. Gitomer, pp. 99–130. Thousand Oaks, CA: Sage Publications.

Harris, Douglas N. 2008b. Breaking the logjam on teacher value-added. *Education Week* 27 (42): 16 June. Available www.edweek.org/ew/articles/2008/06/18/42harris-com_web.h27.html. Accessed 10 May 2008.

Harris, Douglas N. 2009. Toward policy-relevant benchmarks for interpreting effects sizes: Combining effects with costs. *Educational Evaluation and Policy Analysis* 31 (1): 3–29.

Harris, Douglas N. 2010. Education production functions: Concepts. In *International encyclopedia of education,* 3rd ed., edited by Eva Baker, Barry McGaw, and Penelope L. Peterson. Oxford, UK: Elsevier. In press.

Harris, D., and C. Herrington. 2006. Accountability, standards, and the growing achievement gap: Lessons from the past half-century. *American Journal of Education* 112 (2): 209–38.

Harris, Douglas N., and Daniel McCaffrey. 2009. Value-added: Assessing teachers' contributions to student achievement. In *Handbook of teacher assessment and teacher quality,* edited by Mary Kennedy. San Francisco: Jossey Bass. In press.

Harris, Douglas N., and Stacey R. Rutledge. 2010. Models and predictors of teacher effectiveness: A review of the evidence with lessons from (and for) other occupations. *Teachers College Record.* In press.

Harris, Douglas N., and Tim R. Sass. 2005. Value-added models and the measurement of teacher quality. Paper presented at the annual conference of the American Education Finance Association, Louisville, KY, March.

Harris, Douglas N., and Tim R. Sass. 2007a. Teacher training, teacher quality, and student achievement. CALDER Working Paper No. 3, Urban Institute.

Harris, Douglas N., and Tim R. Sass. 2007b. The effects of NBPTS-certified teachers on student achievement. CALDER Working Paper No. 4, Urban Institute.

Harris, Douglas N., and Tim R. Sass. 2007c. What makes a good teacher and who can tell? Paper presented at the summer workshop of the National Bureau of Economic Research, Cambridge, MA, June.

Harris, Douglas N., Lori Taylor, Amy Albee, William K. Ingle, and Leslie McDonald. 2008. The resource cost of standards, assessments and accountability. Paper presented at the National Academy of Sciences Workshop Series on State Standards, Washington, DC, January.

Jacob, Brian A., and Lars Lefgren. 2005. Principals as agents: Subjective performance measurement in education. NBER Working Paper No. 11463.

Kane, Thomas J., and Douglas O. Staiger. 2001. Improving school accountability measures. NBER Working Paper No. 8156.

Kane, Thomas J., and Douglas O. Staiger. 2002. The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives* 16 (4): 91–114.

Kane, Thomas J., and Douglas O. Staiger. 2008. Are teacher-level value-added estimates biased? An experimental validation of non-experimental estimates. Paper presented at the National Conference on Value-Added Modeling, University of Wisconsin–Madison, April.

Koedel, Cory, and Julian R. Betts. 2007. Re-examining the role of teacher quality in the educational production function. Working Paper No. 2007–03, National Center on Performance Initiatives.

Labaree, David. 1997. *How to succeed in school without really learning.* New Haven: Yale University Press.

Lee, Valerie E., and David T. Burkham. 2002. *Inequality at the starting gate.* Washington, DC: Economic Policy Institute.

Levin, Henry M. 1991. Cost-effectiveness at a quarter century. In *Evaluation and education at quarter century*, edited by M. W. McLaughlin and D. C. Phillips, pp. 189–209. Chicago: University of Chicago Press.

Levin, Henry M., and Patrick McEwan. 2001. *Cost-effectiveness analysis*, 2nd ed. London: Sage Publications.

Lockwood, J. R., and Daniel McCaffrey. 2009. Exploring student-teacher interactions in longitudinal achievement data. *Education Finance and Policy* 4 (4): 439–67.

McCaffrey, Daniel, and Laura Hamilton. 2007. Value-added assessment in practice: Lessons from the Pennsylvania value-added assessment system pilot project. Santa Monica, CA: RAND Corporation.

McCaffrey, Daniel, Tim R. Sass, J. R. Lockwood, and Kata Mihaly. 2009. The intertemporal variability of teacher effect estimates. *Education Finance and Policy* 4 (4): 572–606.

Monk, David H. 1987. Assigning elementary pupils to their teachers. *Elementary School Journal* 88 (2): 166–87.

Murnane, Richard J., and David Cohen. 1986. Merit pay and the evaluation problem: Why most merit pay plans fail and a few survive. *Harvard Educational Review* 56: 1–17.

Oakes, Jeannie. 1985. *Keeping track: How schools structure inequality.* New Haven: Yale University Press.

Ogbu, John U. 2003. *Black American students in an affluent suburb: A study of academic disengagement.* Mahwah, NJ: Lawrence Erlbaum.

Podgursky, Michael, and Matthew Springer. 2007. Teacher performance pay: A survey. *Journal of Policy Analysis and Management* 24 (4): 909–49.

Ray, Andrew, Tanya McCormack, and Helen Evans. 2009. Value added in English schools. *Education Finance and Policy* 4 (4): 415–38.

Rice, Jennifer K. 2002. Cost analysis in education policy research: A comparative analysis across fields of public policy. In *Cost-effectiveness in educational policy*, edited by Henry M. Levin and Patrick J. McEwan, pp. 21–35. Larchmont, NY: Eye on Education.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. 2005. Teachers, schools and academic achievement. *Econometrica* 73: 417–58.

Rothstein, Jesse. 2008. Teacher quality in educational production: Tracing, decay, and student achievement. NBER Working Paper No. 14442.

Rothstein, Jesse. 2009. Student sorting and bias in value-added estimation: Selection on observables and unobservables. *Education Finance and Policy* 4 (4): 537–71.

Rothstein, Richard. 2004. *Class and schools: Using social, economic, and educational reform to close the black-white achievement gap*. Washington, DC: Economic Policy Institute; New York: Teacher's College Press.

Sanders, William L., and Sandra P. Horn. 1998. Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education* 12: 247–56.

Toch, Thomas, and Douglas N. Harris. 2008. Salvaging accountability. *Education Week* 1 October. Available www.edweek.org/ew/articles/2008/10/01/06toch_ep.h28.html. Accessed 2 June 2009.

Todd, Petra E., and Kenneth I. Wolpin. 2003. On the specification and estimation of the production function for cognitive achievement. *Economic Journal* 113: F3–F33.